Minireview. Molecular evolution of the neurohypophysial hormone precursors in mammals: Comparative genomics reveals novel mammalian oxytocin and vasopressin analogues

Michael Wallis

Biochemistry Department, School of Life Sciences, University of Sussex, Brighton BN1 9QG. UK.

email: m.wallis@sussex.ac.uk

tel: 01273 472552

Abstract

Among vertebrates the neurohypophysial hormones show considerable variation. However, in eutherian mammals they have been considered rather conserved, with arginine vasopressin (AVP) and oxytocin (OT) in all species except pig and some relatives, where lysine vasopressin replaces AVP. The availability of genomic data for a wide range of mammals makes it possible to assess whether these peptides and their precursors may be more variable in Eutheria than previously suspected. A survey of these data confirms that AVP and OT occur in most eutherians, but with exceptions. In a New-World monkey (marmoset, *Callithrix jacchus*) and in tree shrew (*Tupaia belangeri*), Pro^{8}OT replaces OT, confirming a recent report for these species. In armadillo (*Dasypus novemcinctus*) Leu^{3}OT replaces OT, while in tenrec (*Echinops telfairi*) Thr^{4}AVP replaces AVP. In these two species there is also evidence for additional genes/pseudogenes, encoding much-modified forms of AVP, but in most other eutherian species there is no evidence for additional neurohypophysial hormone genes. Evolutionary analysis shows that sequences of eutherian neurohypophysial hormone precursors are generally strongly conserved, particularly those regions encoding active peptide and neurophysin. The close association between OT and VP genes has led to frequent gene conversion of sequences encoding neurophysins. A monotreme, platypus (*Ornithorhynchus anatinus*) has genes for OT and AVP, organized tail-to-tail as in eutherians, but in marsupials 3-4 genes are present for neurohypophysial hormones, organized tail-to-head as in lower vertebrates.

1. Introduction

The comparative endocrinology of the neurohypophysial hormones has been studied intensively, and a variety of related peptides has been revealed in non-mammalian vertebrates [1,15,29,30]. However, in eutherians, until recently, just vasopressin (VP) and oxytocin (OT) had been reported, the only variation being that in pig and some related cetartiodactyls Lys^8VP (LVP) replaces Arg^8VP (AVP) [10]. Recently it has been shown that Pro^8OT replaces oxytocin in many New World monkeys and in the tree shrew (*Tupaia belangeri*) [20]. In all cases studied genes encoding the precursors of these peptides in Eutheria are adjacent on the same chromosome (chromosome 20 in man), arranged tail-to-tail (i.e. transcribed from opposite DNA strands) [24] and separated by ~10 kilobases (kb) in most species including rat [24, 27], but less in mouse and some other rodents (~3.5 kb) [13].

Marsupials show a more complex picture, with at least 3 neurohypophysial hormones; species studied include the South American opossum (*Didelphis marsupialis*; OT, LVP, AVP [5]), the Eastern Grey Kangaroo (*Macropus giganteus*; mesotocin (MT), LVP and phenypressin [7]) and grey short-tailed opossum (*Monodelphis domestica*; MT, LVP and AVP [11]). In *Monodelphis* these are encoded by 4 genes closely linked on chromosome 5 (arranged: LVP-MT-AVP-MT), transcribed from the same DNA strand [11]. The monotreme platypus (*Ornithorhynchus anatinus*) possesses OT and AVP, like most eutherians [6]. Thus in all mammals for which information is available (and indeed for most lower vertebrates) the genes encoding OT and VP or equivalent peptides are closely linked; this indicates that the two (or more) genes arose by local tandem duplication early in vertebrate evolution rather than by whole genome duplication.

Neurohypophysial peptides are synthesized as precursors, in which a signal peptide is followed by the neurohypophysial peptide and then neurophysin (Np), and in some cases, including VP precursor, an additional peptide, copeptin. Np acts as a binding protein for the conjunct neurohypophysial peptide, though the presence in invertebrates of Np-like proteins that are not associated with VP/OT-like peptides suggest that it may have additional functions [9]. The function of copeptin is unclear. Basic residues between various components of the precursor serve as signals to allow efficient and specific cleavage, and a Gly residue following the neurohypophysial peptide allows C-terminal amidation. The genes for OT and VP each comprise three exons, the first encoding the signal peptide, neurohypophysial peptide and the start of Np, the second the bulk of Np and the third the C-terminal end of Np and (in the case of VP precursor) copeptin [17,19,26].

The availability of genomic sequence data for a large number of mammals provides a basis for more-extensive comparative studies than previously carried out. Here these data have been surveyed to investigate the molecular evolution of the neurohypophysial hormone precursors in mammals. Specific questions addressed include: (1) Does a survey of all the eutherian species for which genomic data are available, indicate the existence of novel neurohypophysial hormones? (2) To what extent do the various functional regions of the precursors evolve at varying rates? (3) How widely does gene conversion between the second exon of VP and OT precursor genes occur? (4) How variable is the overall organization of mammalian neurohypophysial hormone genes ?

2. Mammalian neurohypophysial hormone precursor gene sequences

Complete or partial sequences for the genes encoding the OT and VP precursors were derived from genomic data available for 37 eutherian species, as summarized in Table 1. For most species just one gene was detected for the oxytocin precursor and one for the vasopressin precursor, encoding

normal OT or AVP (LVP in pig) as expected. However, for marmoset (*Callithrix jacchus*), tree shrew, tenrec (*Echinops telfairi*) and armadillo (*Dasypus novemcinctus*), evidence for modified forms of OT or VP was obtained, as is discussed further below. For tenrec and armadillo, additional genes encoding precursors of extensively substituted VP were identified, and in hyrax (*Procavia capensis*) an additional OT precursor gene may occur, though polymorphism at a single locus might also explain the data. Additional genes in some other species cannot be ruled out, but most of the genomic datasets studied are reasonably complete, and it seems unlikely that undetected additional genes would be present, unless very different from those detected, or very similar (differences between genes with very similar sequences may be misinterpreted as polymorphisms if the sequence of a locus is incomplete).

In the marsupial opossum (*Monodelphis domestica*) the presence of four genes was confirmed, encoding MT (two genes), LVP and AVP . Similar paralagous genes were also found in the genome of a second marsupial, wallaby (*Macropus eugenii*), with phenypressin replacing AVP (as in *Macropus giganteus,* [7]), but the gene order was uncertain. In the monotreme platypus genes encoding AVP and OT were identified, as in most eutherians, confirming the observations of Chauvet et al. [6] (see Introduction).

3. Novel eutherian neurohypophysial hormones.

Marmoset and tree shrew. In these two species evidence for an unusual OT-like peptide was obtained - $Pro^{8}OT$, confirming the recent report of Lee et al. [20]. Pro rather than Leu at position 8 would be expected to markedly change the conformation and properties of the peptide, given the difference in side-chain size and the conformational constraints associated with a prolyl residue.

Tenrec. The genomic data for the neurohypophysial hormone precursors in the Madagascan hedgehog/tenrec (*Echinops telfairi*) indicate a complex situation. Three genes, encoding precursors of OT, Thr^4AVP, and a heavily substituted form of vasopressin (sequence: CYLPNCVKD), are well defined (though the sequence data are incomplete for intron 1 of the OT gene), and analysis of data in the Short Read Archive database for *Echinops* confirms the presence of these three sequences. It is notable that conversion of Gln to Thr at position 4 in VP requires base substitutions at both positions 1 and 2 in the codon, suggesting that an intermediate peptide, with Lys or Ala at position 4, occurred in an evolutionary ancestor. Whether the heavily substituted VP is a functioning peptide is not clear. It is not followed in the precursor by Gly, so is probably not amidated, but in other respects the precursor looks functional, with retention of initiating Met, normal splicing signals, basic residues following the vasopressin like sequence, and generally conserved signal peptide and Np (Fig. 1b). In addition to these three well-defined genes, there is evidence for additional genes encoding precursors of Thr^4AVP and a second heavily substituted VP, but complete gene sequences for these could not be constructed. It is not clear whether any of these substantially modified precursors is expressed and processed to give active peptides. No evidence was seen for a precursor of "normal" (Gln^4) AVP, suggesting that Thr^4AVP has replaced it as the functional peptide in tenrec.

Armadillo. The genes for the neurohypophysial peptide precursors for the armadillo indicate that in this species AVP and an unusual form of OT, Leu^3OT occur. Support for the latter is strong, with many sequence reads in the trace archive The available sequence data also suggested the

presence of additional OT/VP related genes in this species, including a much substituted VP precursor gene, probably a pseudogene. No support for "normal" (Ile^3) OT was seen, suggesting that Leu^3OT has replaced this as the expressed OT-like peptide in armadillo.

4. Evolution of Neurohypophysial hormone precursors in eutherians.

Fig. 1a shows an alignment of the available eutherian OT precursors, derived by conceptual translation of the genomic sequences. Sequences are generally well conserved. Notable is the complete conservation of the sequence following OT (GKR), specifying cleavage and amidation [28, 36].

Fig. 1b shows an alignment of the available eutherian VP precursors. Again sequences are conserved, and the GKR sequence following VP is modified in only one case (armadillo), where Arg replaces Lys, a change unlikely to affect cleavage, given the specificity of the converting enzymes [28, 36]. However, in two species, mouse and tenrec, the Arg following Np II is substituted (by His and Leu respectively) which would be expected to affect cleavage of copeptin from Np, which is thought to require a basic side chain [19].

To assess the relative importance of the various regions of the OT and VP precursors, we can investigate their rates of evolution. To do this dN/dS ratios were determined, using the codeml programme [34,35] with a defined tree based on the mammalian phylogeny given in [21]. dN and dS are respectively the rates of evolution at nonsynonymous and synonymous sites within a coding sequence (sites where substitution changes or does not change the amino acid sequence). In most coding sequences $dN << dS$ because purifying selection screens out most substitutions that change the amino acid sequence, whereas synonymous substitutions are largely neutral (without adaptive

effect). If dN/dS = 1, this may indicate a sequence with no function, evolving 'neutrally' without constraint (e.g. a pseudogene). A dN/dS value significantly greater than 1, with nonsynonymous substitutions accumulating more rapidly than 'neutral' synonymous ones, provides clear evidence for adaptive (selective) evolution, though a lower value does not necessarily rule out selection.

An alignment of sequences encoding OT precursors was analysed using codeml. Evidence for some variation of evolutionary rates between species was obtained, but in no case did dN/dS approach 1.0. Constraining dN/dS to a single value for the whole evolutionary tree gave a value of 0.072, indicating that the sequence is strongly conserved. Analysis of separate parts of the precursor showed that the signal peptide shows a notably low dN/dS value of 0.075, much lower than that seen in many other signal sequences (c.f. insulin precursor, dN/dS = 0.32 for primates; TRH precursor, dN/dS = 0.34 for eutherian mammals [32,33]); the signal peptide may be conserved here because it is relatively short, and immediately followed by the biologically active neurohypophysial peptide. dN/dS is very low (0.009) for the sequence encoding OT, reflecting strong conservation. The Np sequence is also strongly conserved (dN/dS = 0.077), suggesting that the role played by Np in binding OT is physiologically important, and requires well-defined structure.

An equivalent analysis of sequences encoding the VP precursor again gave some evidence for variable evolutionary rates. Constraining dN/dS to a single value for the whole evolutionary tree gave a value of 0.092, again indicating a strongly conserved sequence. The signal peptide here is less strongly conserved than for OT (dN/dS = 0.154), though still more so than in many other peptide hormone precursors. dN/dS is very low for vasopressin (0.005), again showing very strong conservation. dN/dS for the sequence encoding copeptin (0.128) is higher than that for Np (0.073; $P<0.01$, likelihood ratio test), but still fairly low, suggesting a specific function for this peptide.

5. Gene conversion

Ruppert et al. [23] reported that the bovine OT and VP genes showed almost complete identity for most of exon 2 (encoding the bulk of Np) and the adjacent part of intron 1, and concluded that this reflected recent gene conversion. They suggested that an equivalent gene conversion had occurred in rat, but at an earlier date, since the converted sequences were less similar. Gwee et al. [11] observed that the second exons of vasotocin (VT) and MT genes in the coelacanth are almost identical, but considered that this was due to purifying selection rather than gene conversion.

To assess the importance of gene conversion in the evolution of eutherian OT and VP genes , the programme GENECONV [25] was used, with an alignment of the first 2 exons of OT and VP, for those species for which complete OT and VP genes were available (exons 3 of OT and VP are very different, and a satisfactory alignment is not possible, and the same is largely true for introns 1 and 2). The results (Table 2) suggested that gene conversion has occurred between exon 2 of OT and VP for most of the species considered. This mostly appears to reflect separate gene conversion events, some rather recent, with the gene converted sequences being identical (e.g. bovine, guinea pig), but others less recent, with accumulation of a few point mutations since the gene conversion event (e.g. rat, armadillo). Visual inspection showed that in some cases the gene conversion event was confined to exon 2, while in others it extended into the last part of intron 1 (e.g. bovine) or the first part of intron 2 (e.g. guinea pig). The only species for which such gene conversion was not evident were the bush baby and the Old-World monkeys macaque and baboon; the conversion in other primates was relatively 'weak' (i.e. less recent), with only a relatively short sequence identified (chimpanzee), or a fair number of mismatches (man, orangutan, gibbon and marmoset).

The consequence of such gene conversion is that the sequences of Nps I and II encoded by exon 2 are more similar in any one species than would be expected, and often more similar than are the sequences of Np I (or II) from related species. This suggests that although the sequence of Np is strongly conserved, reflecting functional constraints including binding of OT and VP, specificity with regard to the two hormones (if there is any) may lie in the N- or C-terminal ends encoded by exons 1 or 3, so that conversion of most of Np I to Np II, or *vice versa*, could occur without affecting functionality, possibly promoting frequent gene conversion. Gene conversion between exons 1 is prevented, however, because it would often involve loss of one or other of the biologically active neurohypophysial peptides. Gene conversion between exons 3 is ruled out because exons 3 of OT and VP are so different that the homology required for effective gene conversion is no longer present.

6. Organization of neurohypophysial hormone genes

Gwee et al. [11] described the organization of the neurohypophysial hormone genes in the opossum (*Monodelphis domestica)*, showing that it differs markedly from that in eutherians and resembles that in many non-mammalian vertebrates in that all four genes are arranged tail-to-head, i.e. transcribed from the same DNA strand (Fig.2; see [11, 12] for a summary of the organization of neurohypophysial hormone genes in lower vertebrates). This was interpreted as indicating that a genomic reorganization occurred on the lineage leading to eutherians, giving the tail-to-tail organization of OT and VP genes.

Examination of genome sequences confirmed the organization proposed in [11] for opossum and showed that a second marsupial, wallaby (*Macropus*), also has 2 genes encoding MT (though one of these lacks a start codon at the normal position and therefore may not be expressed), together with

genes encoding LVP and phenypressin. However, the monotreme platypus has only genes encoding AVP and OT, in accord with the earlier report of Chauvet et al. [6]. In the genomic assembly these are organized tail-to-tail as in eutherians, and examination of mate pairs for the sequencing traces confirmed the tail-to-tail arrangement, with the OT and AVP genes in platypus separated by about 8 kb. It is generally accepted that monotremes diverged from the therian line before separation of marsupials and eutherians (e.g. [4, 21]), implying that either the genomic reorganization leading to the tail-to-tail arrangement of the OT and VP genes occurred twice, independently on lineages leading to monotremes and eutherians, or that a single reorganization of this sort occurred early in mammalian evolution but was then reversed during marsupial evolution.

Whether the tail-to-tail arrangement of OT and VP genes occurs in all eutherians is not yet firmly determined, given the fragmentary nature of many of the genomic assemblies, but this arrangement applies in every case in which the data is sufficiently clear-cut to reach a conclusion, including several primates, rodents and artiodactyls, and dog (Fig. 2). Elucidation of the arrangement of the multiple OT/VP genes in armadillo and tenrec will have to await more complete assemblies.

7. Conclusions

The genomic resources currently available provide a rich source of comparative information, especially for eutherian mammals. The data are in many cases rather incomplete, but the survey carried out here has provided a number of clear conclusions, together with pointers for further work.

1. It is clear that the diversity of neurohypophysial hormones seen in Eutheria is greater than has previously been appreciated, with novel peptides having arisen on at least five independent occasions (Fig. 2). In both tenrec and armadillo, not only are novel neurohypophysial hormones

produced, but there is clear evidence for more than one VP-encoding gene. In this respect it is notable that in many species there are multiple forms of VP receptor [8]; multiple VP genes might be expressed in different tissues, serving different receptors. In most eutherian species, however, there is no evidence for additional genes for neurohypophysial hormone precursors, and no support for reports of vasotocin in pineal gland from man or other eutherians [3,22].

2. The gene conversion that was noted previously [23] for bovine and rat OT- and VP-encoding genes occurs very widely. It is confined mainly to exon 2, encoding the bulk of the Np sequence, and sometimes extends for a short way into upstream or downstream intron sequence. As a consequence in many species the sequences encoding Np I and Np II (including synonymous sites) are very similar, much more similar than are the orthologous sequences in quite closely related species, despite the ancient origin of the two genes by tandem duplication. It thus appears that in this region repeated gene conversion events throughout the long period since gene duplication (more than 400 million years) have offset the expected effects of divergent evolution.

3. The observation that the organization of neurohypophysial hormone genes in platypus resembles that in eutherians (tail-to-tail arrangement of OT and VP encoding genes) rather than that in marsupials and lower vertebrates (tail-to-head arrangement), clearly has implications for the distinct organization of the VP and OT genes seen in Eutheria. The data suggest a surprising evolutionary scenario, involving either independent reorganization on lineages leading to both monotremes and eutherians to give the tail-to-tail arrangement, or reversion from tail-to-tail to tail-to-head arrangement during evolution of marsupials (Fig. 2).

References

[1] R. Acher, Neurohypophysial peptide systems: processing machinery, hydroosmotic regulation, adaptation and evolution, Reg. Peptides 45 (1993) 1-13.

[2] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic Local Alignment Search Tool, J. Mol. Biol. 215 (1990) 403–410.

[3] C. Badiu, M. Coculescu, M. Møller, Arginine vasotocin mRNA revealed by in situ hybridization in bovine pineal gland cells, Cell Tissue Res. 295 (1999) 225-229.

[4] O.R.P. Bininda-Emonds, M. Cardillo, K.E. Jones, R.D.E. MacPhee, R.M.D. Beck, R. Grenyer, S.A.Price, R.A. Vos, J.L. Gittleman, A. Purvis, The delayed rise of present-day mammals, Nature 446 (2007) 507-512.

[5] J. Chauvet, D. Hurpet, M.T. Chauvet, R. Acher, Divergent neuropeptide evolutionary drifts between American and Australian marsupials, Bioscience Reports 4 (1984) 245-252.

[6] J. Chauvet, D. Hurpet, G. Michel, M.T. Chauvet, F.N. Carrick, R. Acher, The neurohypophysial hormones of the egg-laying mammals: identification of arginine vasopressin in the platypus (*Ornithorhynchus anatinus*), Biochem. Biophys. Res. Commun. 127 (1985) 277-282.

[7] M.T. Chauvet, D. Hurpet, J. Chauvet, R. Acher, Identification of mesotocin, lysine vasopressin, and phenypressin in the eastern gray kangaroo (*Macropus giganteus*), Gen. Comp. Endocrinol. 49 (1983) 63-72.

[8] D.O. Daza, M. Lewicka, D. Larhammar, The oxytocin/vasopressin receptor family has at least five members in the gnathostome lineage, including two distinct V2 subtypes, Gen Comp. Endocrinol. 175 (2012) 135-143.

[9] M.R. Elphick, NG peptides: A novel family of neurophysin-associated neuropeptides, Gene 458 (2010) 20-26.

[10] D.R. Ferguson, Genetic distribution of vasopressins in the peccary (*Tayassu angulatus*) and warthog (*Phacochoerus aethiopicus*), Gen. Comp. Endocrinol. 12 (1969) 609-613.

[11] P-C. Gwee, C.T. Amemiya, S. Brenner, B. Venkatesh, Sequence and organization of coelacanth neurohypophysial hormone genes: evolutionary history of the vertebrate neurohypophysial hormone gene locus, BMC Evo. Biol. 8 (2008) 93.

[12] P-C. Gwee, B-H. Tay, S. Brenner, B. Venkatesh, Characterization of the neurohypophysial hormone gene loci in elephant shark and the Japanese lamprey: origin of the vertebrate neurohypophysial hormone genes, BMC Evo. Biol. 9 (2009) 47.

[13] Y. Hara, J. Battey, H. Gainer, Structure of mouse vasopressin and oxytocin genes, Mol. Brain Res. 8 (1990) 319-324.

[14] D.G. Higgins, P.M. Sharp, CLUSTAL: a package for performing multiple sequence alignment on a microcomputer, Gene 73 (1988) 237-244.

[15] C.H.V. Hoyle, Neuropeptide families: evolutionary perspectives, Reg.Peptides 73 (1998) 1-33.

[16] T.J.P. Hubbard, B.L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, et al., Ensembl 2007, Nucl. Acids Res. 35 (2007) D610-D617.

[17] R. Ivell, D. Richter, Structure and comparison of the oxytocin and vasopressin genes from rat, Proc. Natl. Acad. Sci. USA. 81 (1984) 2006-2010.

[18] W.J. Kent, BLAT - the BLAST-like alignment tool, Genome Res. 12 (2002) 656-664.

[19] H. Land, G. Schütz, H. Schmale, D. Richter, Nucleotide sequence of the cloned cDNA encoding bovine arginine vasopressin-neurophysin II precursor, Nature 295 (1982) 299-303.

[20] A.G. Lee, D.R. Cool, W.C. Grunwald, D.E. Neal, C.L. Buckmaster, M.Y. Cheng, S.A. Hyde, D.M. Lyons, K.J. Parker, A novel form of oxytocin in New World monkeys, Biol. Lett. 7 (2011) 584-587.

[21] W.J. Murphy, P.A. Pevzner, S.J. O'Brien, Mammalian phylogenomics comes of age, TRENDS in Genetics 20 (2004) 631-639.

[22] S. Pavel, M. Dorcescu, R. Petrescu-Holban, E. Ghinea, Biosynthesis of a vasotocin-like peptide in cell cultures from pineal glands of human fetuses, Science 181 (1973) 1252–1253.

[23] S. Ruppert, G. Scherer, G. Schütz, Recent gene conversion involving bovine vasopressin and oxytocin precursor genes suggested by nucleotide sequence, Nature 308 (1984) 554-557.

[24] E. Sausville, D. Carney, J. Battey, The human vasopressin gene is linked to the oxytocin gene and is selectively expressed in a cultured lung cancer cell line, J. Biol. Chem. 260 (1985) 10236-10241.

[25] S.A. Sawyer, GENECONV: a computer package for the statistical detection of gene conversion (1999) Distributed by the author, Department of Mathematics, Washington University, St. Louis, MO; available at: http://www.math.wustl.edu/~sawyer

[26] H. Schmale, S. Heinsohn, D. Richter, Structural organization of the rat gene for the arginine vasopressin-neurophysin precursor, EMBO J. 2 (1983) 763-767.

[27] E. Schmitz, E. Mohr, D. Richter, Rat vasopressin and oxytocin genes are linked by a long interspersed repeated DNA element (LINE) - sequence and transcriptional analysis of LINE, DNA Cell Biol. 10 (1991) 81-91.

[28] N. G. Seidah, M. Chrétien, Proprotein and prohormone convertases: a family of subtilases generating diverse bioactive polypeptides, Brain Res. 848 (1999) 45–62.

[29] A. Urano, H. Ando, Diversity of the hypothalamo-neurohypophysial system and its hormonal genes, Gen. Comp. Endocrinol. 170 (2011) 41–56.

[30] M. Wallis, Molecular evolution of pituitary hormones, Biol. Rev. 50 (1975) 35-98.

[31] M. Wallis, Mammalian genome projects reveal new growth hormone (GH) sequences. Characterization of the GH-encoding genes of armadillo (*Dasypus novemcinctus*), hedgehog (*Erinaceus europaeus*), bat (*Myotis lucifugus*), hyrax (*Procavia capensis*), shrew (*Sorex araneus*), ground squirrel (*Spermophilus tridecemlineatus*), elephant (*Loxodonta africana*), cat (*Felis catus*) and opossum (*Monodelphis domestica*), Gen. Comp. Endocrinol. 155 (2008) 271-279.

[32] M. Wallis, New insulin-like growth factor (IGF)-precursor sequences from mammalian genomes: the molecular evolution of IGFs and associated peptides in primates, GH & IGF Res. 19 (2009) 12-23.

[33] M. Wallis, Molecular evolution of the thyrotrophin-releasing hormone precursor in vertebrates: Insights from comparative genomics, J. Neuroendocrinol. 22 (2010) 608-619.

[34] Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood, Mol. Biol. Evol. 24 (2007) 1586–1591.

[35] Z. Yang, R. Nielsen, Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages, Mol. Biol. Evol. 19 (2002) 908–917.

[36] A. Zhou, G. Webb, X. Zhu, D. F. Steiner, Proteolytic processing in the secretory pathway, J. Biol. Chem. 274 (1999) 20745–20748.

Legends for figures.

Fig.1. Alignment of precursor sequences for oxytocin (a) and vasopressin (b) . In each case the sequence of the human protein is shown in full; the other sequences are compared with this, with . representing identity, and – indicating a gap. Sequences were derived from genomic sequence data, obtained from the ensembl website (release 63, *http://www.ensembl.org*) for species for which genomic assemblies were available [16]. Data from the ncbi trace database (*http://www.ncbi.nlm.nih.gov/Traces*) were used to check and complete these genomic sequences, and in cases where assemblies were not available. Traces were assembled using the Staden Package (*https://sourceforge.net/projects/staden/*). For each species studied, these databases were searched using the BLAST or BLAT search methods [2,18] with the coding sequences (CDS) for human or bovine precursors of OT or AVP, or of other species as these became available. Additional sequence data were also obtained by searching the embl and Uniprot websites. In some cases sequence data were also available on the SRA database (http://trace.ncbi.nlm.nih.gov/Traces/sra) and were used to confirm or extend critical sequences. The quality and accuracy of the sequences used were assessed as in previous studies [31-33], by examining the available traces in detail and by comparison with previously available data, usually protein or cDNA sequence. Doubts about the validity of derived sequences are noted in Table 1, as are places where sequences were incomplete. Details of the mammalian posterior pituitary hormone gene sequences derived in this study are given in Supplementary Figs. 1 and 2. Neurohypophysial hormone precursor gene, CDS and protein sequences were aligned using the Clustalw programme [14], followed by manual adjustment as necessary.

Fig.2. Diagram illustrating the distribution and organization of genes encoding neurohypophysial hormones in relation to the vertebrate phylogenetic tree. The tree is based on [4,21]; branch lengths are not proportional to evolutionary time. Not all species for which

genomic data are available are included. Data for non-mammalian vertebrates are based on [11, 12]. For other teleosts, most are similar to the example shown (puffer fish), but in zebra fish (*Danio rerio)* IT and VT genes occur on separate chromosomes. In all other cases, the genes for neurohypophysial hormones are thought to be adjacent on the same chromosome, in the orientation shown, but for some eutherians the genomic data are too incomplete to establish this unequivocally (shown by the absence of a line between the adjacent genes). For tenrec and armadillo the (incomplete) genomic data suggest that there are additional genes/pseudogenes (see text). Marmoset reflects the position in some other, but not all, New World Monkeys [20]. Unusual mammalian neurohypophysial peptides are highlighted. Scientific names not included in the text or Table 1 are: chicken, *Gallus gallus;* zebra finch, *Taeniopygia guttata*; Xenopus *Xenopus tropicalis*; coelacanth *Latimeria chalumnae;* puffer fish (fugu) *Takifugu rubripes*; elephant shark (a holocephalian) *Callorhinchus milii*; lamprey *Lethenteron japonicum*.

Fig. 1

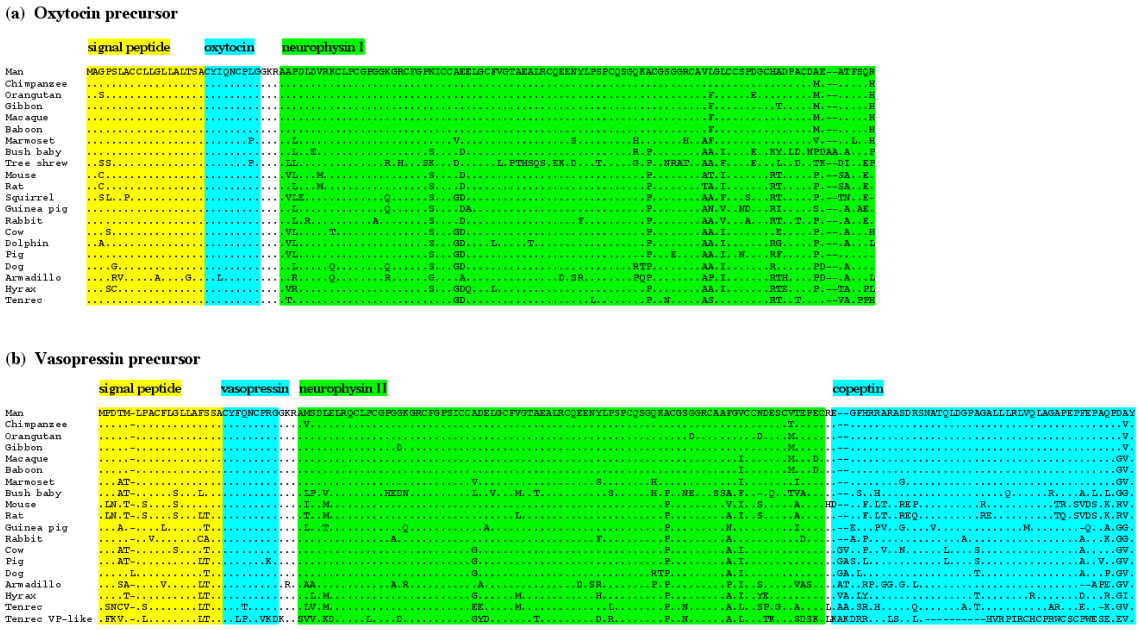
(a) Oxytocin precursor
signal peptide
oxytocin
neurophysin I
Man
Chimpanzee
Orangutan
Gibbon
Macaque
Baboon
Marmoset
Bush baby
Tree shrew
Mouse
Rat
Squirrel
Guinea pig
Rabbit
Cow
Dolphin
Pig
Dog
Armadillo
Hyrax
Tenrec
(b) Vasopressin precursor
signal peptide
vasopressin
neurophysin II
copeptin
Man
Chimpanzee
Orangutan
Gibbon
Macaque
Baboon
Marmoset
Bush baby
Mouse
Rat
Guinea pig
Rabbit
Cow
Pig
Dog
Armadillo
Hyrax
Tenrec
Tenrec VP-like

Fig. 1

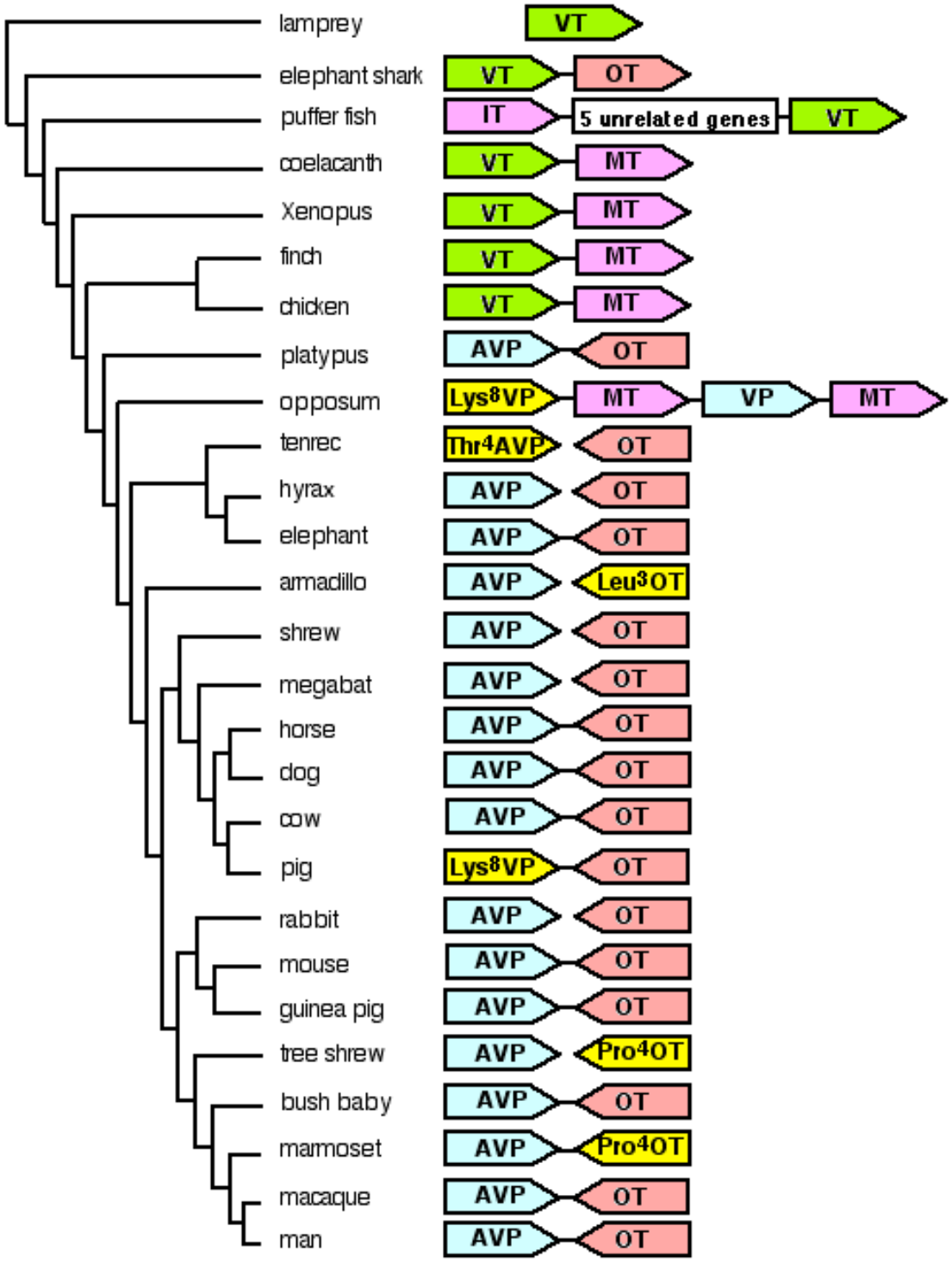
lamprey
VT
elephant shark
VT
OT
puffer fish
IT
5 unrelated genes
VT
coelacanth
VT
MT
Xenopus
VT
MT
finch
VT
MT
chicken
VT
MT
platypus
AVP
OT
opposum
Lys8VP
MT
VP
MT
tenrec
Thr4AVP
OT
hyrax
AVP
OT
elephant
AVP
OT
armadillo
AVP
Leu3OT
shrew
AVP
OT
megabat
AVP
OT
horse
AVP
OT
dog
AVP
OT
cow
AVP
OT
pig
Lys8VP
OT
rabbit
AVP
OT
mouse
AVP
OT
guinea pig
AVP
OT
tree shrew
AVP
Pro4OT
bush baby
AVP
OT
marmoset
AVP
Pro4OT
macaque
AVP
OT
man
AVP
OT

Table 1

Neurohypophysial hormone precursor genes in mammalian genomes as available from genomic data

Species	Common name	OT	AVP	Np I	Np II/ copeptin	Comments	Data source*
Homo sapiens	man	√	√	√	√		a,c,d,e,f,g,m
Pan troglodytes	chimpanzee	√	√	√	√		a,b
Gorilla gorilla	gorilla	√	√	√	√ **	exon 1 for OT ***	c
Pongo pygmaeus	orangutan	√	√	√	√	intron 2 for VP ***	b
Nomascus leucogenys	gibbon	√	√	√	√ (intron 2 ***)		b,d
Macaca mulatta	macaque	√	√	√	√		b,d,e
Papio hamadryas	baboon	√	√	√	√		d
Callithrix jacchus	marmoset	√ Pro^8OT	√	√	√	intron 1 for OT ***	b,d
Tarsius syrichta	tarsier	√	√	n/a	n/a		b
Microcebus murinus	mouse lemur	√	√	√ **	n/a		a
Otolemur garnettii	bush baby	√	√	√	√		a
Tupaia belangeri	tree shrew	√ Pro^8OT	√	√	√ **	intron 1 for OT ***	a
Mus musculus	mouse	√	√	√	√		a,b,c,f
Rattus norvegicus	rat	√	√	√	√		c,f,g,h
Spermophilus tridecemlineatus	squirrel	√	√	√	√ **		a
Dipodomys ordii	kangaroo rat	√	√	n/a	√ **		d
Cavia porcellus	guinea pig	√	√	√	√		a
Oryctolagus cuniculus	rabbit	√	√	√	√		a
Ochonta princeps	pika	√ ***	n/a	n/a	√ **		a
Bos taurus	cow	√	√	√	√		d
Ovis aries	sheep	√	√	√	√	from mRNA sequences	n
Tursiops truncatus	dolphin	√	√	√	√ **		d
Vicugna pacos	alpaca	n/a	√	n/a	n/a		b
Sus scrofa	pig	√	√ LVP	√	√		c,i,j
Canis familiaris	dog	√	√	√	√		a,k
Felis catus	cat	√	√	√ **	mostly missing		g
Ailuropoda melanoluca	panda	n/a	√	n/a	√ **		o
Equus caballus	horse	√	√	√ **	√(exon 3***)		a
Myotis lucifugus	microbat	n/a	√	?	?	2 Np sequences **	a
Pteropus vampyrus	megabat	√	√	√ **	√ **		d
Erinaceus europeaus	hedgehog	n/a	√	√ **	n/a		a
Sorex araneus	shrew	√	√	n/a	√ **		a
Dasypus novemcinctus	armadillo	√ Leu^3OT	√	√	√	at least one additional VP-type gene/pseudogene	a,d
Loxodonta africana	elephant	√ ***	√ ***	n/a	√		a
Procavia capensis	hyrax	√	√	√	√	possibly 2 OT genes	d
Echinops telfairi	tenrec	√	√ Thr^4VP	√	√	several VP-like genes. OT intron 1 incomplete	a
Choleopus hoffmanni	sloth	n/a	n/a	√? **	n/a		b
Monodelphis domestica	opossum	√ MT	√ LVP,AVP	√	√	4 genes; not all complete	a
Macropus eugenii	wallaby	√ MT	√ LVP, phenypressin	√	n/a	4 genes; one MT gene may not be functional	d,l
Ornithorhynchus anatinus	platypus	√	√	√	√ (gap in intron 1)		b

* Sources of original data: a) The Broad Institute; b) Washington University, Genome Sequencing Center; c) The Sanger Center; d) Baylor College of Medicine; e) J. Craig Venter Institute; f) Celera Genomics; g) Agencourt Bioscience Corporation; h) University of Utah Genome Center; i) National Livestock Research Institute (Korea); j) Sino-Danish Joint Venture Partnership; k) The Institute for Genome Research; l) Australian Genome Research Facility Ltd; m) Cold Spring Harbor Laboratory; n) Genbank/EMBL/DDJ entries X16052 (OT) and EU045357 (VP); o) Beijing Genomics Institute.
√ indicates that the sequence was obtained; n/a indicates that a complete sequence was not available (but should NOT be interpreted as absent in this species). ** incomplete sequence; *** poor sequence.

Table 2
Gene conversion between oxytocin-like and vasopressin-like genes

Sequences	Base positions	Length (bases)	No of mismatches	p value
Cow OT:VP	123-318	196	0	<0.0001
Guinea pig OT:VP	122-324	203	0	<0.0001
Rabbit OT:VP	122-318	197	0	<0.0001
Chimpanzee OT:VP	217-299	83	0	0.0015
Rat OT:VP	142-299	158	1	<0.0001
Dog OT:VP	121-318	198	1	<0.0001
Mouse OT:VP	122-303	182	3	<0.0001
Hyrax OT:VP	122-318	197	4	<0.0001
Armadillo OT:VP	121-321	201	5	<0.0001
Man OT:VP	122-299	178	6	<0.0001
Pig OT:VP	122-321	200	6	0.0018
Orangutan OT:VP	122-286	165	6	0.0020
Tenrec OT:VP	136-324	189	7	<0.0001
Gibbon OT:VP	122-299	178	7	0.0030
Marmoset OT:VP	122-292	171	10	0.0394

Gene conversion was studied using GENECONV [23], run using an alignment (324 bases) of sequences for exons 1 and 2 (exon 2 starting at base 121) and with some mismatches allowed (setting g1). 'Base positions' indicate the sequence location for which gene conversion (reflecting significant sequence similarity) was detected. p values are corrected for multiple hypothesis testing. Grouping was used to restrict analysis to within-species events.